LBL-12216
Preprint

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

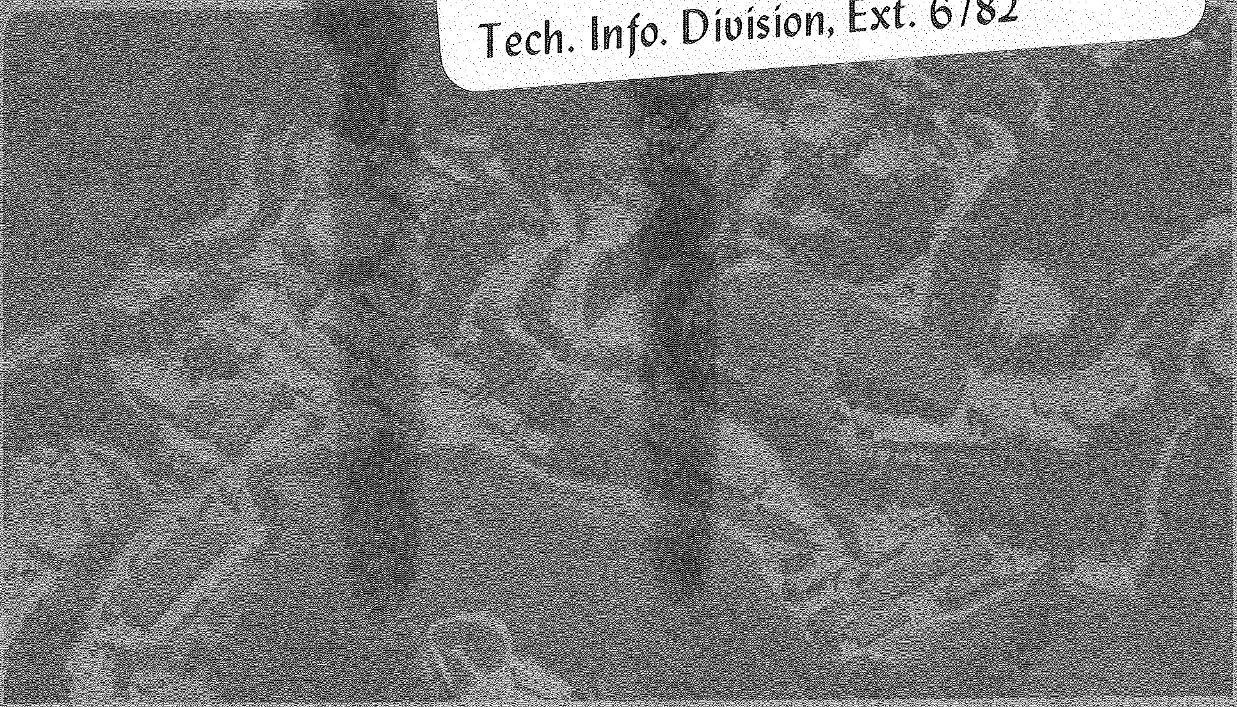
Physics, Computer Science &
Mathematics Division

Submitted to the American Journal of Epidemiology

AN ALTERNATIVE TO ECOLOGIC REGRESSION ANALYSIS OF MORTALITY RATES

S. Selvin, D. Merrill, and S.T. Sacks

February 1981



Prepared for the U.S. Department of Energy under Contract W-7405-ENG-48

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

AN ALTERNATIVE TO ECOLOGIC REGRESSION ANALYSIS OF MORTALITY RATES

S. Selvin[1], D. Merrill[2] and S. T. Sacks[3]

Lawrence Berkeley Laboratory
University of California
Berkeley, CA 94720

1. Department of Biomedical and Environmental Health Sciences, University of California, Berkeley, Berkeley, CA 94720.
2. Department of Computer Science and Mathematics, Lawrence Berkeley Laboratory, Berkeley, CA 94720.
3. Department of Epidemiology and International Health, University of California, San Francisco, San Francisco, CA 94143.

The work described in this paper was funded by the Office of Health and Environmental Research, Assistant Secretary for Environment of the U.S. Department of Energy under Contract No. W-7405-ENG-48.

ABSTRACT

Selvin, S. (Department of Biomedical and Environmental Health Sciences, University of California, Berkeley, CA 94720), D. Merrill and S. T. Sacks. AM J Epidemiol xxx:xxx-xxx, 1981.

A number of recent papers use geographically defined data and linear models to study the relationship between a series of epidemiologic factors and the frequency of disease. This "ecologic regression" approach involves serious problems of interpretation. An alternate approach is discussed that does not depend on statistical models, produces easily interpreted results, and yields statistical summaries that approximately parallel regression analysis. This alternative procedure is illustrated with a small set of national leukemia mortality data.

Analysis of Mortality Rates

Basic to epidemiology is the identification and assessment of factors influencing disease. One analytic approach involves linear models with disease rates serving as the dependent variable and a series of independent variables derived from summaries of aggregated data (sometimes called ecologic regression analysis). For example, the application of regression techniques to geographic units rather than individuals is the basis for several studies of the relationship between mortality and air pollution [e.g., references (1) or (2)]. The possible utility and potential pitfalls ("ecologic fallacies") of analyzing ecologic data became focused when Robinson (3) pointed out that the behavior of a series of individuals cannot be usually inferred from the analysis of summary values associated with aggregated units. Although Robinson's remarks dealt with the application of product-moment correlation coefficients and other authors have continued to describe the problems of ecologic inferences [e.g., references (4), (5), or (6)], little has been offered in the way of approaches to analyzing mortality data.

Increased availability of computer technology has lead to greater use of large data files, particularly nationwide geographically based data. For example, recently published data involving the more than 3000 U.S. counties, used in conjunction with cardiovascular disease mortality rates (7), illustrates the use of ecologic regression analysis in epidemiology. The extensive cancer mortality maps published by the National Cancer Institute [(8), (9)] are further examples of the use of geographically based data for investigating disease etiology. Geographic variation in leukemia rates among U.S. counties has also been quantitatively analyzed by investigators at National Cancer Institute using regression models applied to age-adjusted mortality rates (10). The following discussion presents an analytic strategy for investigating disease rates and their relationship to large sets of ecologic data. This approach parallels to some extent ecologic regression analysis without some of the well documented problems. The statistical approach to be presented applies to a wide range of situations. However, for the sake of concreteness and to avoid general terminology, the discussion will be in terms of mortality rates analyzed in the context of county level data.

County mortality data are often summarized by correlation coefficients (7) or regression coefficients (10). Several problems are of immediate concern when mortality rates are treated with regression techniques. Mortality rates are generally not normally distributed, which implies that statistical tests lead to, at best, approximate significance probabilities. County level mortality rates also differ widely in precision (variance) due to the large differences in county population. This fact adds another disrupting factor in statistical analyses although weighted analyses can be performed. Models exist for transforming rates to produce dependent variables that more closely conform to the structure required for valid statistical analysis. However, transformed rates are rather artificial values with little intuitive appeal and are usually no longer independent of the influences of population size, which is the reason for calculating rates in the first place. In addition to statistical difficulties, the epidemiological interpretation of regression or correlation coefficients is especially complicated when ecologic data are employed. The statistical issues and problems of interpreting ecologic regression analysis are fully discussed elsewhere [e.g., references (4) and (11)]. Of course, the adequacy of a linear model is always a concern for any regression analysis regardless of the properties of the dependent and independent variables. In spite of the problems associated with ecologic data analysis these data are readily available, usually relatively inexpensive to obtain, and can cover the entire U.S. on at least a county level for a large number of variables.

A simple linear regression coefficient or correlation coefficient is often computed from a set of data to determine the extent to which two variables are linearly related. More generally, two variables are associated when ordering one induces in the other some non-random pattern (not necessarily linear). This fact can be used as a basis of investigating disease rates and their associations to ecologic variables. Clearly, this approach to defining an association is suited only for variables that can be ordered. Acute lymphocytic leukemia mortality among white female children less than 5 years of age during the period 1969-1977 (12) will be used to illustrate various points throughout the discussion -- Table 1 gives some brief summary statistics.

Suppose interest is focused on the existence of a relationship between leukemia and socio-economic status (SES). The percentage of county residents earning more than $15,000 per year (in 1969) is one of the many variables available from U.S. Census data and will serve to relate childhood leukemia to at least one dimension of socio-economic status. A computer file of data is constructed, which contains for each county 1) percent income ⩾ $15,000 2) the number of deaths from acute lymphocytic leukemia among white females under 5 years old, and 3) an estimate of the person-years at risk. In general terms, the file contains 1) an independent or predictor variable, 2) a dependent variable, and 3) an estimate of person-years at risk for each geographic unit. This file is ranked from low to high on the basis of the income variable. The number of deaths are then accumulated into a series of group with exactly equal numbers of person-years at risk so that the file now consists of a series of equal-risk groups and the county nature of the data is no longer relevant. Deaths to residents of counties not entirely included in a single group are proportionally divided among the categories overlapped by that county. (A FORTRAN listing of a subroutine that performs this task is available from the authors.)

An illustration of ten equal-risk categories, is given in Table 2. The number of counties (as defined by the Johns Hopkins Mortality Surveillance Program) is 3075. Six counties have no estimate for percent earning ⩾ $15,000 and are discarded from the sample. In the remaining 3069 counties, during the nine years 1969-1977, 605 deaths from acute lymphatic leukemia were reported among white females less than 5 years old. (There were no deaths in the six discarded counties.) The corresponding estimate of persons-at-risk is 59,961,008, which is the sum of the estimated populations (for white females under 5 years) for each of the nine years 1969, 1970, ... 1977. (In this calculation the 1969 population, which was not available, was assumed equal to the 1970 population.) The 605 deaths are distributed into 10 equal-risk categories each containing 5,996,100.8 person-years of risk (illustrated in table 2). In the following analyses 200 rather than 10 equal-risk categories were used where the population-at-risk in each of the 200 equal-risk categories is 59,961,008/200 = 299,805.04.

If the predictor variable is unrelated to the frequency of mortality, then the expected number of deaths in each equal-risk group is simply estimated by the overall mean value. Under the null hypothesis that an independent or predictor variable (for example, percent of persons having income ⩾ $15,000) is stochastically independent of the dependent variable (for example, acute lymphocytic leukemia mortality), then the expected number of deaths in each of 200 equal-risk categories is estimated by the mean (e.g., $\bar{x}$ = 605/200 = 3.02 deaths). Furthermore, under the hypothesis that the predictor variable is unrelated to the dependent variable, the number of deaths per category follows a Poisson or Binomial distribution. Note that the Poisson property results when no relationship exists between the predictor and dependent variable, and should not be confused with the occasionally made assumption that the number of deaths from a rare disease follows a Poisson distribution. The number of deaths from any cause or, in fact, any discrete variable that has a small, constant and independent probability of falling into one of a series of categories will follow a Poisson distribution. Therefore, the null hypothesis of no association between a disease entity and a predictor variable is readily quantified and statistically tested. It should be noted that the expected number of deaths (3.02) in each equal-risk category is equal to the product of the mortality rate (1.009 per 100,000) times the population-at-risk (59,961,008 / 200 = 299,805.04) in each category.

The null hypothesis of no association can be tested by contrasting the observed variation (S^2) in the number of deaths with the expected variance from a Poisson distributed variable which is estimated by the mean value ($\bar{x}$). The ratio of these two estimated variances ($S^2/\bar{x}$) should be near 1.0 when no association exists between the predictor variable and the number of deaths or the mortality rate. If the predictor variable is related to the frequency of disease, then an increase over the expected variance should be observed. The common chi-square test ($\Sigma(obs_i - ex_i)^2/ex_i$ where $ex_i = \bar{x}$) and the chi-square test of variance ($(k-1)S^2/\bar{x}$ where k = number of equal-risk categories) are the same in this case and easily lead to significance probabilities ("p-values"). For example, the observed variation in the number of female leukemia deaths among the 200 equal-risk categories ordered by income (% ⩾ $15,000) is 3.23. This gives a ratio of $S^2/\bar{x} = 1.07$, yielding a chi-square statistic of 212.9 with a significance probability of 0.238 (table 3).

The use of a chi-square statistic applied to data grouped into equal-risk categories is conservative. That is, if it is unlikely no association exists (small significance probability) using aggregated data, then it is more unlikely that no association exists in the ungrouped (but unavailable) data. This is not the case when regression coefficients are employed to test for an association between two ecologic variables. The aggregation bias [defined in (4)] of ecologic regression coefficients can either increase or decrease the observed value. A statistical test of these regression coefficient, in most cases, will also be conservative (e.g., understate the t-test statistic) but not always.

The observed variation in the number of deaths can be standardized to a number between 0 and 1.0 to create a summary value analogous to the squared multiple correlation coefficient used in regression analysis (i.e., R^2, the percentage of the variation "explained"). The maximum variance in the number of deaths among a series of equal-risk categories occurs when the dependent variable is ordered before the data is aggregated into equal-size groups.

Another view of this relationship comes from noticing that when the number of deaths serves as both the predictor and the dependent variables, the variability among categories (predictability) is maximized. In the case of female childhood leukemia the maximum possible variability among the 200 categories is 12.99, resulting in a "correlation-coefficient-like" value of "R^2" = 3.23/12.99 = 0.249 associated with income (% ≥ $15,000).

In order to implement the strategy of employing equal-risk groups, a choice must be made for the number of groups. If only a few groups are chosen, the breadth and variability of the data are lost. If too many groups are chosen, it then defeats the purpose of grouping the data. Experience with a number of nationwide data sets shows that the observed variance (S^2) decreases slightly as the number of groups chosen increases. However, empirically it seems that, for county level data, the ratio $S^2/\bar{x}$ is more or less stable for a number of groups between 100 and 500. Neverthèless, relative comparison of these ratios is useful even when there is some subjectivity in the actual statistical test procedure. This same phenomenon is usually an issue with most chi-square tests employing categorical data.

Any ordinal variable can be used as the predictor in this approach using equal-risk categories to analyze mortality data. Several papers in the literature discuss urban/rural influences on leukemia rates [e.g., references (10) or (13)]. Employing the percentage of urban area in each county (percent urban) as a predictor variable produces an observed variation in the number of deaths among the 200 equal-risk categories of $S^2 = 3.43$ (Table 3). Since the expected variance remains equal to the mean $\bar{x} = 3.02$ deaths per category, the ratio $S^2/\bar{x} = 1.14$ yields a chi-square statistic of 226.8 with a "p-value" = 0.082 (table 3). Similarly, elevation above sea level is occasionally discussed as a risk factor associated with leukemia (14). In this case, the predictor variable of elevation induces an empirical variance of $S^2 = 3.45$ and, again, compared to an expected value of 3.02 yields a "p-value" = 0.078 (Table 3). In both cases, "R^2" = 0.27, which indicates that both predictor variables (percent urban and elevation) are moderately related to the frequency of childhood acute lymphocytic leukemia nationwide.

Computer mapping has recently become a popular method of studying nationwide mortality rates [e.g., (7), (8) and (9)]. Often, these maps can be somewhat difficult to interpret objectively. For example, the large western counties have a disproportionate visual impact when a map of the entire U.S. is considered. The partitioning of mortality data into equal-risk categories with respect to a specific predictor variable is an effective analytic tool for assessing geographic patterns. When the predictor variables are the county centroid latitude and longitude or functions thereof, the observed variation in mortality can be used to evaluate the strength of geographic patterns. Sorting numbers of deaths into a series of equal-risk categories based on, say, latitude (or longitude) is equivalent to counting the numbers of deaths in a series of geographic strips having equal populations-at-risk. If the observed variability does not differ significantly from the variance expected under the Poisson assumption, then it is reasonable to conclude that the data do not reflect a detectable geographic patterns. Rejecting the Poisson hypothesis, on the other hand, implies that geographic patterns are likely to exist. The data on childhood leukemia show a strong association with latitude ($S^2 = 4.48$ implying $p < 0.001$ table 3) and none with longitude ($S^2 = 3.13$ implying $p = 0.355$ table 3). It is also possible that more complicated patterns are of interest. For example, one might hypothesize from the results of other analyses that low rates are seen in the center of the U.S. and high rates near the boundaries. This possibility can be assessed using as a predictor variable an index that is low near the geographic center of the country and increases toward the borders. The childhood lymphocytic leukemia analyzed with such an index (index = $(\text{longitude} - 91.9)^2 + (\text{latitude} - 38.3)^2$ where (91.9, 38.3) is the approximate position (in degrees) of the U.S. geographic centroid) yields an observed variation in the number of deaths of $S^2 = 4.34$ ($p < 0.001$), indicating that a two-dimensional parabolic function significantly describes the leukemia mortality pattern.

An important feature of regression analysis is the possibility to add or remove variables from the analysis and observe the influence on some measure of change such as the residual sum of squares. Similarly, an index, say the first principal component, can be formed; then, changes in the variation (S^2) of the number of deaths resulting from the addition or subtraction of predictor variables from this index can serve to indicate the relative influence of subsets of predictor variables. For example, the first principal component based on the variance-covariance array of the ten ecologic variables (percent migration, percent urban, percent black, percent earning, ≤ \$3,000, percent earning ≥ \$15,000, percent professional, percent employed in manufacturing, percent college educated, percent foreign residents, and elevation) produces an observed variation in leukemia deaths of $S^2 = 3.22$ ("R^2" = 0.248) among the 200 equal-risk categories (table 3). If all but the two income variables (percent ≤ \$3,000 and percent ≥ \$15,000) are removed from the analysis and the first principal component based on just these two variables is used as a predictor, the observed variation is only slightly reduced, to $S^2 = 3.19$ ("R^2" = 0.246). This inconsequential change indicates that the major influence among the ten predictor variables is income-related. The use of a canonical index in conjunction with the equal-risk procedure is analogous to the "extra sum of squares" principle (15) employed in multiple regression analysis, but lacks a formal "F-to-remove" test. Also, each variable added to a linear regression analysis monotonically increases the regression sum of squares. This is not the case for an index based on principal components since the observed variation S^2 can either increase or decrease when variables are added to the index. However, the adding and removing of variables from an index to assess the multivariate impact on S^2 is essentially assumption-free. The suggestion to use an index to provide a multivariate approach to the analysis of leukemia mortality is one of many possibilities. Other, perhaps more sensitive, techniques could indeed be employed and their statistical properties examined.

The classification of data into equal-risk categories is not automatically superior or inferior to "ecologic regression analysis." For example, a regression analysis produces a predictive equation that can be used to make estimates of mortality rates which are potentially useful and are not produced by using equal-risk categories. Like any analytic technique, the properties underlying the data will dictate the appropriate analysis. The suggested approaches, which are by no means exhaustive, merely indicate possible ways of dealing with some problems inherent in county-level mortality data to test for epidemiologic relevant associations.

Analysis of Mortality Rates

1. Lave L, Seskin E. Air Pollution and Human Health. Baltimore, MD: Johns Hopkins Press, 1977.

2. Mendelsohn R, Orcutt G. An empirical analysis of air pollution dose-response curves. J Envir Econ and Management 1979; 6:85-106.

3. Robinson WS. Ecological correlations and the behavior of individuals. Am Sociol Rev. 1950; 15:351-357.

4. Langbein LI, Lichtman AJ. Ecologic Inference. Beverly Hills, CA: Sage Publications, 1978.

5. Kasl SV. Mortality and the business cycle: Some questions about research strategies when utilizing macrosocial and ecological data. Am J Public Health 1979; 69:784-788.

6. Goodman L. Some alternatives to ecologic correlation. Am J Sociol 1959; 64:610-625.

7. Fabsitz R, Feinleib M. Geographic patterns in county mortality rates from cardiovascular disease. Am J Epidemiol 1980; 111:315-328.

8. Mason TJ, McKay FW, Hoover R, et al. Atlas of Cancer Mortality for U.S. Counties: 1950-1969. Washington, DC, U.S. Department of Health, Education and Welfare Pub No (NIH) 75-780.

9. Mason TJ, McKay FW, Hoover R, et al. Atlas of Cancer Mortality for U.S. Counties: 1950-1969. Washington, DC, U.S. Department of Health, Education and Welfare Pub No (NIH) 76-1204.

10. Blair A, Fraumeni JF, Mason TJ. Geographic patterns of leukemia in the United States. J Chron Dis 1980; 33:251-260.

11. Gorgatta EF, Jackson FJ. Aggregated Data Analysis and Interpretation. Beverly Hills, CA: Sage Publications, 1980.

12. Gittelsohn A. Data tabulated by the "Mortality Surveillance Project", Johns Hopkins University, Baltimore, MD, 1980.

13. MacMahan B. Geographic variation in leukemia mortality in the United States. Pub Hlth Rep 1957; 72:39-46.

14. Eckhoff ND, Shultz RW, Clark RW, Ramer ER. Correlation of leukemia mortality rates with altitude in the United States. Health Physics 1974; 27:377-380.

15. Draper NR, Smith H: Applied Regression Analysis. New York, NY: John Wiley & Sons, 1966.

Table 1

Summary statistics for acute lymphocytic leukemia U.S. mortality* among white females under five years old (1969-1977).

Deaths in county	Number of counties	Total deaths	Population at risk
0	2699	0	28,831,365
1	254	254	11,510,954
2	69	138	6,796,916
3	23	69	3,376,622
4+	24	144	9,445,151
Total	3069	605	59,961,008

average annual mortality = 1.009/100,000.

maximum deaths in county = 29 (Los Angeles County, CA)

*Data source: National Center for Health Statistics, tabulated by Alan Gittelshohn, Johns Hopkins University (12).

Table 2

Sample calculation for 10 equal risk categories using U.S. mortality per 100,000 from acute lymphocytic leukemia for white females age 0-4 (1969-1977)

Sample	Number of Counties	Population-at-Risk	Percent ≥ $15,000	Deaths 1969-1977	Annual Mortality
01	1213.1	5996100.8	6.4	59.0	.98
02	702.2	5996100.8	10.4	61.0	1.02
03	443.0	5996100.8	13.3	66.0	1.10
04	265.6	5996100.8	16.0	65.0	1.08
05	140.5	5996100.8	18.4	70.9	1.18
06	102.9	5996100.8	20.6	64.8	1.08
07	84.2	5996100.8	23.4	54.4	.91
08	53.0	5996100.8	27.1	62.1	1.04
09	24.8	5996100.8	30.3	48.5	.81
10	39.7	5996100.8	38.7	53.3	.89
TOTAL	3069.0	59961008.0	20.5	605.0	1.01
MISSING	6.0	17384.0			

Table 3

Various statistical summaries relating acute lymphocytic leukemia mortality among white females under five years old (1969-1977) to a series of predictor variables.

Predictor	S^2	$S^2/\bar{x}$	"p-value"	"R^2"
Income (% ⩾ $15,000)	3.23	1.07	0.238	0.249
Urban (% urban)	3.43	1.14	0.082	0.265
Elevation	3.45	1.15	0.078	0.266
Geographic				
Longitude	3.13	1.04	0.335	0.241
Latitude	4.48	1.49	<0.001	0.345
Index	4.34	1.44	<0.001	0.334
Principal Component Index				
Index -- all 10 variables	3.22	1.07	0.238	0.248
Index -- two income variables	3.19	1.06	0.268	0.246